DOCUMENT RESUME

ED 123 251 TM 005 310

AUTHOR Hynes, Kevin
TITLE Cross-Validation, Shrinkage, and Multiple Regression.
PUB DATE [Apr 76]
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Correlation; Error Patterns; Literature Reviews; *Mathematical Models; *Multiple Regression Analysis; Prediction; *Research Methodology; Statistical Analysis; *Validity
IDENTIFIERS Cross Validation; *Shrinkage (Multiple Correlation)

ABSTRACT

One aspect of multiple regression--the shrinkage of the multiple correlation coefficient on cross-validation is reviewed. The paper consists of four sections. In section one, the distinction between a fixed and a random multiple regression model is made explicit. In section two, the cross-validation paradigm and an explanation for the occurrence of shrinkage are presented. In section three, research on shrinkage by Burket (1964) and by Brown (1969) is covered. Finally, in section four, shrinkage literature is considered with respect to theoretical-methodological advances and the development of better computer programs. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SESSION 13.25

ED123251

Cross-Validation, Shrinkage, and Multiple Regression[1]

Kevin Hynes

Purdue University

The purpose of this paper is to review one aspect of multiple regression--the shrinkage of the multiple correlation coefficient upon cross-validation. The paper consists of four sections. In section one, the distinction between a fixed and a random multiple regression model is made explicit. In section two, the cross-validation paradigm and an explanation for the occurrence of shrinkage are presented. In section three, research on shrinkage by Burket (1964) and by Brown (1969) is covered. Finally, in section four, shrinkage literature is considered with respect to theoretical-methodological advances and the development of better computer programs.

---

[1]A paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1976. The author wishes to thank Dr. Hubert E. Brogden for his guidance throughout the preparation of this paper. The interpretations reported here, however, are solely the responsibility of the author.

TM005 310

Fixed vs. Random Predictor Model

Brogden (1972) suggested that fixed models generally are inappropriate for observational or correlational research and that random models generally are inappropriate for controlled experimentation. Since much of the research related to testing and prediction is observational, it follows that a random model, not a fixed model, is appropriate for testing and prediction.

The main reason a random model is appropriate for observational research is that the independent variable in this situation usually is not a manipulable or controllable variable. For example, a test score cannot be manipulated in the same way that variety of seed can be manipulated. Plots of ground can be randomly assigned to variety of seed, but individuals cannot be randomly assigned to test scores. Tests and measures in this sense are really dependent variables.

If the analysis of variance design for a controlled experiment is converted to a regression model, each regression weight can be interpreted as an experimental effect, because the factors are quantifiable and because extraneous factors have been controlled or randomized. However, this is not true for observational data. For observational data, randomization procedures are often lacking. This lack of randomization may result in dependent or correlated errors. Thus, when random sampling of persons and occasions is absent, observational data cannot be treated as a fixed regression model. On the other hand, a random regression model is usually appropriate for observational studies.

Other researchers have also advocated the appropriateness of the random regression model for observational studies. Browne (1969), for example, studied the precision of prediction in the random regression model. He concluded that "the fixed model is more easily handled mathematically but appears to be of less practical use" (p. 89). Browne's findings cast some doubt on Burket's (1964) suggestion that "in prediction problems in psychology, the predictor variables are generally random rather than fixed . . . . However, since correlation theory [random model] is considerably more complex and difficult to apply than regression theory [fixed model], the latter is generally used, with the hope that the practical differences between conclusions drawn from the two models will be negligible." (p. 2).

Moreover, Herzberg (1969) suggested that "a more realistic model for most multivariate work in psychology is to assume [that] both the predictors and the criterion are random variables" (p. 2). Herzberg (1969) stressed that, for the random model, it is important to note whether the constant is included in the regression whenever the mean squared error of prediction is reported because the predictors are random, but the constant term is fixed.

Claudy (1972) maintained that the "misapplication of the Fixed-X model to Random-X data causes an over-fitting of the regression surface to the sample data . . . . This over-fitting, or error fitting, results in two types of errors . . . ." (p. 313). These two types of error will be discussed in the next section of this paper. First, however, the statistical models for the fixed and the random regression models will be considered.

Winer (1971) presents the statistical derivations for the fixed and the random regression models. The derivations begin by postulating a population of elements for which the following model holds:

$$Y = \beta_o + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

or $$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad \text{(matrix form)}$$

For the fixed regression model, the $X_1,\ldots,X_p$ predictors are fixed, the criterion is random, and the error is assumed to be an independently and identically distributed random variable with expectation of zero. Given a random sample of n elements from the population, the least-squares principle can then be used to estimate the parameters, β and σ_ε^2. The least-squares principle minimizes the sum of the squared errors. This means that the sum of the squared deviations of the observed from the predicted criterion scores is minimized.

Then $\underline{b} = (X'X)^{-1}X'Y$ (where X'Y is singular)

and $$\sigma_\varepsilon^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} = \frac{Y'Y - \underline{b}'X'\underline{Y}}{n-p-1} .$$

For the random regression model, the $X_1,\ldots,X_p$ predictors and the criterion (Y) are random variables and are assumed to have a (p+1)-variate normal distribution with parameters $\underline{\mu}$ and Σ. For the random case, $\underline{\beta} = \Sigma_{xx}^{-1}\underline{\sigma}_{xy}$, where Σxx is the population variance-covariance matrix for the p predictors and $\underline{\sigma}xy$ is a vector for the covariances between the predictors and the criterion. The error of prediction is minimized by

$$\sigma_\varepsilon^2 = \sigma^2_{y|\underline{x}} = \sigma^2_y\,(1-\rho^2_{y\cdot 12\ldots p})$$

Given a random sample of n elements from N_{p+1} $(\underline{\mu},\Sigma)$ and using maximum likelihood procedures, th n $\hat{\underline{\beta}} = \underline{b} = V_{xx}^{-1} \underline{v}_{xy}$, where V_{xx} is the sample variance-covariance matrix for the p predictors and $\underline{v}_{xy}$ is a vector of covariances between each of the predictors and the criterion. Also, the population multiple correlation is estimated by

$$\hat{\rho}_{y \cdot 12 \ldots p} = [\hat{\underline{\beta}}' \underline{v}_{xy} / s^2_{(y)}]^{1/2} = \left(\frac{ss_{\hat{Y}}}{ss_Y}\right)^{1/2}$$

where $s^2_{(y)}$ is the sample variance of the criterion.

Since the rational and statistical distinctions between the fixed and the random regression models have been made, the cross-validation paradigm will be described and the occurrence of shrinkage will be explained. The next section will also make a distinction between shrinkage as the expected value of a multiple correlation in a second sample and shrinkage as the loss of efficiency of prediction in a second sample as measured by the mean squared error.

Cross-Validation and Shrinkage

Mosier (1951) described six possible cross-validation designs. The designs were called: cross-validation, validity-generalization, validity extension, simultaneous validation, replication, and double cross-validation. In the cross-validation design the effectiveness of a regression equation is determined by using the regression equation developed in a calibration sample to predict the criterion scores in a separate, independent sample, where both samples are randomly selected from the same universe. Since this paper emphasizes the cross-validation design, the additional designs that Mosier (1951) described will not be considered.

Darlington (1968) provided a clear-cut explanation of the statistical terminology involved in cross-validation. To begin with, three correlations need to be distinguished. The first correlation is the population multiple correlation ($\rho_{y\hat{y}}$). $\rho_{y\hat{y}}$ is the correlation between the criterion scores in the population and the regression function calculated in the population. But since the population regression function is usually unknown, attention shifts to sample statistics. The second correlation, the sample multiple correlation (R), is the Pearson product-moment correlation between the criterion scores in the calibration sample and the regression function calculated in the calibration sample. The third correlation, the cross-validity ($\rho_{yy'}$), is the correlation between the criterion scores in the cross-validation sample and the predicted criterion scores that result from the application of the calibration sample regression function to the cross-validation sample data. The three correlations are estimates of the validity of their respective regression functions.

Another measure of the validity of a regression function is the mean squared error of prediction (MSE). MSE is the average of the squared differences between each person's criterion score and the regression function's prediction of that score.

Darlington (1968) described the mean squared errors that parallel the correlations just presented. In this paper they will be called the population mean squared error, $\sigma^2(\rho_{y\hat{y}})$; the calibration sample mean squared error, $\sigma^2(R)$; and the cross-validation sample mean squared error, $\sigma^2(\rho_{yy'})$. These mean squared errors pertain to the regressions defined by $\rho_{y\hat{y}}$, R, and $\rho_{yy'}$, respectively.

Now that the basic terminology regarding cross-validation has been defined, the occurrence of shrinkage will be explained.

When the multiple correlation is involved in a cross-validation, there are two sources of error that produce shrinkage. Herzberg (1969) explained one reason for the occurrence of shrinkage: "Due to the finite size of the sample, the optimizing linear combination will be fitted to the idiosyncracies of the sample and will generally result in a higher multiple correlation than the population multiple correlation" (p. 1). Therefore, the sample multiple correlation will overestimate the population multiple because of capitalization on chance.

In describing the second reason for shrinkage, Claudy (1972) noted that "the interpredictor variability of the n sample beta weights is inflated, relative to that of the population beta weights. The variability referred to here is not the variance or standard error of a single beta weight for a single predictor variable, but rather the standard deviation of a set of n beta weights derived from a sample of n predictor variables" (p. 313). Since estimates of the beta weights are unstable if the sample is small, shrinkage is greater for small samples. As mentioned earlier, Claudy (1972) suggested that it is the misapplication of the fixed regression model to random variable data that produces both sources of error.

Several researchers (Brogden, 1954; Browne, 1969; Darlington, 1968; Nicholson, 1948) have noted that the mean squared error of a calibration regression function, applied to a cross-validation sample, cannot be converted into an exact estimate of the cross-validity.

The problem is that, upon cross-validation, the criterion is estimated more precisely at the mean and less precisely at high and low criterion values. The result is that a third source of error is introduced when the mean squared error is used to estimate the accuracy of a cross-validation.

This third source of error is graphically demonstrated in Figure 1. The actual criterion values in a cross-validation sample (y) are shown plotted against the predicted criterion values (y') based on the regression function determined in a calibration sample. Two regression lines are considered. If we assumed that the y' were the same as in the calibration sample, then we would expect a 45° regression line (line 1 in Figure 1) because the y' would equal the y mean arrays. In going to a cross-validation sample, however, shrinkage occurs, and the result is a shallower slope (line 2 in Figure 1). The result is that high values of y are overestimated and low values of y are underestimated.

Brogden (1954) explained the problem shown in Figure 1 in this way:

> Consider the bivariate surface formed by plotting, in the second sample, the estimates of y derived from the first sample regression equation against the actual y values. The estimates of y would ordinarily be expected to equal the means of the arrays of y (in a scatter plot based on the sample in which the regression equation was derived). In the bivariate surface in the second sample, the estimates of y no longer equal the means of the arrays of y values. The estimates are biased and consistently overestimate for high positive values of y and underestimate for values of y below the mean (p. 382).

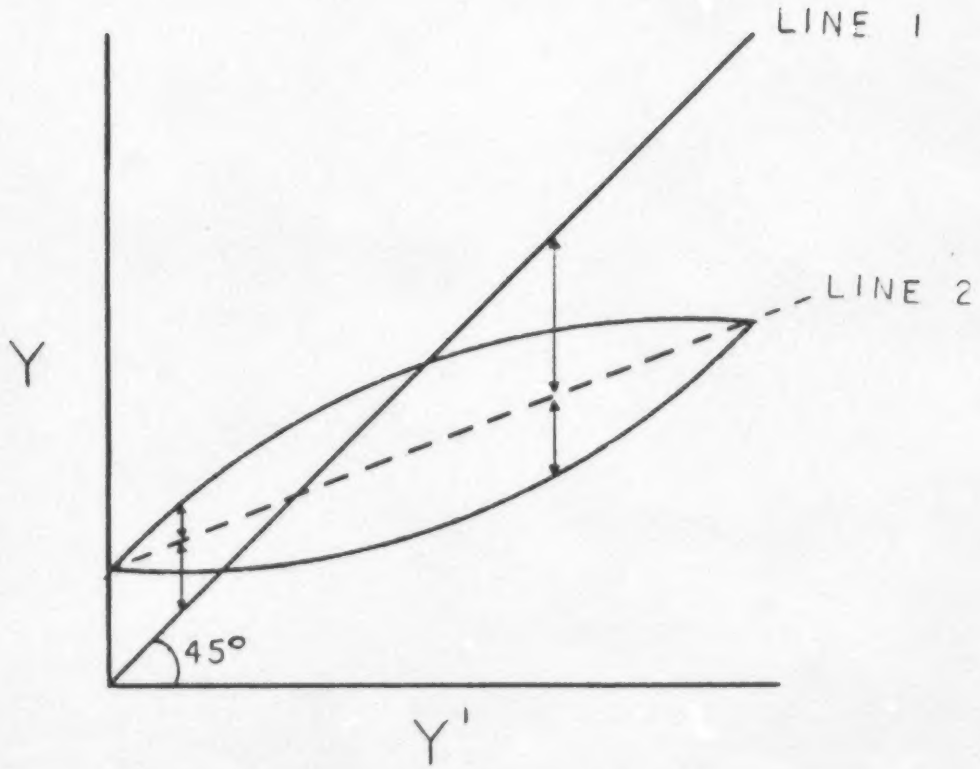

Figure 1. Cross-validation and mean squared error.

Thus, when the multiple correlation is involved in a cross-validation, two sources of error account for shrinkage. First, the sample multiple correlation overestimates the population multiple correlation. Second, the cross-validity underestimates the population multiple correlation. The three correlations are related as follows:

$$R > \rho_{y\hat{y}} > \rho_{yy'}.$$

The sample multiple correlation (R) will overestimate the population multiple correlation (ρ_{yy}), and the cross-validity ($\rho_{yy'}$) will underestimate the population multiple correlation ($\rho_{y\hat{y}}$). However, when the mean squared error is used to estimate the accuracy of a cross-validation, a third source of error is introduced, as explained in the discussion of Figure 1.

Brogden (1954) noted that cross-validation generally implies a concern for predicting only the relative, not the absolute values of the criterion. Moreover, Darlington (1968) noted that "since the correlation coefficient likewise depends only on relative scores, it follows that, in this situation, the correlation between the measure and the criterion variable gives a more realistic statement of the value of the measure than does the mean square error" (p. 175). It is therefore advisable to rely on the multiple correlation rather than the mean squared error, in order to determine the accuracy of a cross-validation.

Shrinkage Literature

In this section, work by Burket (1964) and by Browne (1969) will be considered. Their work has resulted in important advances in estimating shrinkage. Additional references on the shrinkage problem, that will not be considered here, are Nicholson (1948), Lord (1950), Darlington (1968), and Herzberg (1969). The distinction of a fixed versus a random regression model is emphasized in reviewing the articles.

Burket (1964) studied five reduced-rank procedures for regression weights to see if these methods were more accurate than the full-rank model in predicting future samples. Although Burket's work was limited to the fixed regression situation, he proposed that the statistics he developed could be considered estimates for the random regression case. Burket (1964, see Equation 64) proposed the following formula as an estimate of the cross-validity or weight-validity:

$$\rho_{yy'} = \hat{W} = \frac{NR^2 - p}{R(N-p)} = \frac{\hat{\rho}_{y\hat{y}}^2}{R} \qquad (1)$$

The weight-validity, W, will be less than the estimate of the population multiple correlation, $\rho_{y\hat{y}}$.

Burket (1964) also defined $\psi_{\bar{\beta}}$ as "an inverse index of weight-efficiency: the smaller it is, the more suitable $\bar{\beta}$ will be for a prediction problem . . . .$\psi_{\bar{\beta}}$ is not a mathematical expectation in the usual sense, but rather an expectation of an expectation" (p. 11). An unbiased estimate of $\psi_{\hat{\beta}}$ is as follows:

$$\sigma^2(\rho_{yy'}) = \psi_{\hat{\beta}} = \frac{N+p}{N-p}(1-R^2) \qquad (2)$$

It may be useful to provide an example (Brogden, 1975) that demonstrates the use of formulas 1 and 2. Given that sample size, N = 50, and that the number of predictors, p = 10, and that $R^2 = .36$, then

$$\sigma^2(\rho_{yy'}) = \psi_{\hat{\beta}} = \frac{50 + 10}{50 - 10}(1 - .36) = .96$$

Thus, for this example, when error is the estimate of accuracy, it would be pointless to use the regression weights -- the raw data are just as useful. This is because the variables have been normalized so that $\sigma^2(\rho_{yy'})$ equals one by chance alone. Applying formula 1 to the same example, we find that

$$\hat{\rho}_{yy'} = \frac{50\ (.36) - 10}{(.60)(50-10)} = .33.$$

The five reduced-rank procedures that Burket (1964) studied were the following: predictor elimination, predictor selection by accretion, largest principal-axes factors, smallest principal-axes factors, and principal-axes factors giving the highest multiple correlation. The first two procedures (stepwise selection procedures) involve using a subset of the original predictors, and the last three involve using principal-axes factors.

Burket (1964) conducted an empirical comparison of the five reduced-rank procedures. The study involved 29 measures of academic success as the predictors and five grade-point averages as the criteria. The criteria actually served to define five populations. From each population, six random samples (with replacement) ranging from 30 to 255 cases were drawn. For each sample, regression weights were computed using the five rank-reduction procedures for ranks 1 through

29. Cross-validation was accomplished by randomly sampling from cases remaining in the population after the calibration sample had been selected. The calibration weights were then applied to the cross-validation sample, and the weight validity was estimated. Cross-validation sample size was always 252 cases.

Results showed that the method of smallest principal-axes factors was substantially worse than the other methods. The method of largest principal-axes factors was generally superior to the other rank-reduction procedures, especially for small sample sizes. For the larger sample sizes, other procedures were occasionally better than the largest principal-axes factor method. On the basis of these results, Burket (1964) suggested that the method of largest principal-axes factor might be useful in two situations--in order to maximize accuracy of prediction for a given calibration sample size or in order to minimize calibration sample size for a given accuracy of prediction. However, $\hat{\psi}_\beta$ was not found to be a satisfactory basis for selection.

The work done by Browne (1969), emphasized the random regression model. The squared weight validity, w^2, and the predictive mean squared error, d^2, were the two measures of accuracy used in the study. A direct comparison between the random and the fixed regression models was made to determine the effect of using a fixed regression model in order to analyze data involving random variables.

Browne (1969) presents the following estimate of mean squared error:

$$d^2 = \frac{(n-2)(n+1)}{(n-p)(n-p)} s_e^2$$

If n = 50, p = 10, and R^2 = .36, then d^2 = .979. This estimate of the mean squared error of prediction is larger than the estimate obtained when the Burket formula was applied. When the Burket formula was applied, $\sigma^2(\rho_{yy'})$ was equal to .96. As an estimate of cross-validity, Browne (1969) proposed the following:

$$w^2 = \frac{(n-p-2)\hat{\rho}^4 + \hat{\rho}^2}{(n-p)\,\hat{\rho}^2 + (p-1)}$$

where

$$\hat{\rho}^4 = \max \left[(\hat{\rho}^2) - \frac{2(p-1)(1-\hat{\rho}^2)}{(n-1)(n-p)} \right]^2$$

Browne (1969) found that the fixed and the random regression models differ substantially when N-p is small, but do not differ much when N-p is large. Larger variances of d^2 were found for the random regression model. Browne also presented a comparison of the squared weight validity, w^2, for the random and the fixed predictor models. Browne (1969) reported that "the difference in bias of r^2 between the fixed predictor model and the random predictor model is small" (p. 83). Thus the use of a random, as opposed to a fixed, predictor model becomes important in estimating d^2 or w^2 with a small sample and many predictors. But if the sample is large, the two models produce very similar results. Browne (1969) reported that Burket's estimator is an adequate estimate of w^2 when N-p is moderately large.

Summary and Implications for the Educational Researcher

In this section, the practical implications of shrinkage as a methodological concern of the educational researcher will be considered. Ideally, such computer programs as the regression subroutine of the Statistical Package for the Social Sciences (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) would provide options both for the selection of a fixed or a random regression solution and for the selection of several ways to estimate shrinkage. Currently, the SPSS regression subroutine permits only a fixed regression solution and provides the following adjustment to the multiple correlation (Nie, et al., 1975, p. 358):

$$R^2(\text{adjusted}) = R^2 - \frac{(p-1)}{(n-p)} \quad (1-R^2)$$

where p is the number of predictors and n is the sample size. Some versions of SPSS do not provide the adjusted R^2 figure. How, then, should the educational researcher deal with shrinkage?

Three suggestions are presented as one way to deal with shrinkage. First, be sure that the ratio of sample size to the number of predictors is large. Studies with a sample size less than 100 may wish to consider either using simple unit predictor weights (i.e., simple summing of standardized test scores) or using the techniques developed by Browne (1969).

Second, estimate the cross-validity, $\rho_{yy'}$, using a computer program designed for a fixed regression. In the SPSS package, this can be accomplished by first obtaining a regression function based

on the calibration sample. The beta weights from this calibration regression function can then be applied to the cross-validation sample in order to obtain the predicted y-values. Within the SPSS package, this can be accomplished using the COMPUTE procedure, where predicted y is the sum of the calibration regression weights times the predictors in the cross-validation sample. The researcher should realize that estimating cross-validity for random-X data using a fixed regression program is justifiable only when the ratio of sample size to predictors is large. Browne (1969) demonstrated that the fixed and the random regression models yield comparable results when the ratio of sample size to predictors is large.

Third, using the sample multiple correlation from step one, apply the following formula (Burket, 1964) to obtain an estimate of cross-validity:

$$\rho_{yy'} = \hat{W} = \frac{NR^2 - p}{R(n-p)}$$

where R^2 is from step one, n is the sample size, and p is the number of predictors. The result should be similar to the result from step two. Following these three suggestions should allow the educational researcher to obtain a cross-validity estimate that correctly accounts for shrinkage.

References

Brogden, H. E. Personal communication, April, 1975.

Brogden, H. E. Some observations on two methods in psychology. Psychological Bulletin, 1972, 77, 431-437.

Brogden, H. E. Statistical theory and research design. Annual Review of Psychology, 1954, 5, 377-400.

Browne, M. W. Precision of prediction. Research Bulletin 69-69. Princeton, N. J.: Educational Testing Service, 1969.

Burket, G. R. A study of reduced rank models for multiple prediction. Psychometric Monograph, 1964, No. 12.

Claudy, J. G. A comparison of five variable weighting procedures. Educational and Psychological Measurement, 1972, 32, 311-322.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Herzberg, P. A. The parameters of cross-validation. Psychometrika (Monograph Supplement 16), 1969, 34.

Lord, F. M. Efficiency of prediction when a regression equation from one sample is used in a new sample. Research Bulletin 50-40. Princeton, N. J.: Educational Testing Service, 1950.

Mosier, C. I. Problems and designs of cross validation. Educational and Psychological Measurement, 1951, 11, 5-11.

Nicholson, G. E. The application of a regression sample to a new sample. Unpublished doctoral dissertation, University of North Carolina, 1948.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.